School absenteeism and Dropout - A Baysian duration model

Sofie J. Cabus†† Kristof De Witte§

January 25, 2012

Abstract

School absenteeism (or truancy) may be a signal of an ongoing process of student attrition that eventually leads to early school leaving. This paper estimates by a duration model the influence of school absenteeism on the risk of dropout in secondary education. In particular, using a semiparametric Cox duration model, the timing of the dropout decision of truants is compared with the timing of regular school attendees. Using administrative data with insights on wrongly specified truancy spells, we correct in a Bayesian duration model for issues on data uncertainty. We show that only a Bayesian structure on the data yields insightful and consistent results. The analysis provides evidence of poor data validity on post-compulsory secondary school students aged 18 years old or older. Moreover, the results indicate that school-aged students enrolled in vocational education are most vulnerable to unauthorized school absenteeism and dropout. A relative risk ratio of 1.374 significant at 1%-level is estimated, indicating that truants are 37.4 percent more likely to dropout of school than regular school attendees.

Keywords: Bayesian duration model; School absenteeism; Early school leaving

Jel: I21, I28

---

††We are grateful to Wim Groot, Henriëtte Maassen van den Brink and TIER seminar participants for useful comments and discussions. We would like to thank Orm Muilwijk of DMO Amsterdam and the Dutch Ministry of Education for providing the data. The authors acknowledge financial support of NICIS. The usual caveat applies.

Top institute for Evidence Based Education Research, Maastricht University, Kapoenstraat 2, 6200 MD Maastricht, the Netherlands, s.cabus@maastrichtuniversity.nl.

Corresponding author.

§Top institute for Evidence Based Education Research, Maastricht University, Kapoenstraat 2, 6200 MD Maastricht, the Netherlands, k.dewitte@maastrichtuniversity.nl and Faculty of Economics and Business, University of Leuven (KUL), Naamsestraat 69, 3000 Leuven, Belgium, kristof.dewitte@econ.kuleuven.be.
1 Introduction

Students who leave secondary education without a school-leaving certificate have only bleak prospects.\(^1\)\(^2\) At various levels, we observe undesired pecuniary and non-pecuniary outcomes associated with early school-leaving. We summarize them at three levels: (1) private costs; (2) costs to society; and (3) loss in fiscal revenues. Private costs denote higher risks on, for example, long-term unemployment (e.g., Psacharopoulos and Layard, 1979), or lower health status (e.g., Kaplan\textit{ et al.}, 1994; Groot and Maassen van den Brink, 2007). Subsequently, loss in fiscal revenues arise from lower tax revenues, and higher unemployment and/or health insurance benefits (Psacharopoulos, 2007). Furthermore, school dropout is positively associated with poverty (Bowles, 1972; Milligan\textit{ et al.}, 2004), juvenile crime (Lochner and Moretti, 2004) and a slowdown in economic growth (Hanushek and Wößmann, 2007). This, in turn, imposes increased costs to society.

The European Council decided at the 2000 Lisbon Summit to aim for a reduction of the rate of early school-leaving at secondary education from 20\% to 10\% by the year 2010. This 10 percentage point reduction has not been reached by 2010 as the average EU early school leaving rate amounts to 15.6\% (Eurostat). The ambition has therefore been recently restated in the Europe 2020 flagship initiative ‘Youth on the Move’, where a target of 8\% youth leaving school without a higher secondary degree is advocated. Within Europe 2020, early school leaving is even considered as a pivot of obtaining other headline targets: employment, R&D and innovation, climate change and energy, and poverty (European Commission, 2010). The European Commission (2011, p.2) states that "Drastically reducing the numbers of young people leaving school early is a key investment not only in the prospects of each and every one of its young people but also in the future prosperity and social cohesion of the EU in general." A similar political agenda is maintained overseas, where the president of the United States Barack Obama (2010) states in its Reauthorization of the Elementary and Secondary Education Act that school dropout is no longer an option.

It has been extensively discussed in many different ways that the most effective policies to reduce early school leaving are preventive measures (for some examples, see: Adams and Becker, 1990; Jencks and Mayer, 1990; Bell\textit{ et al.}, 1994; Rumberger and Larson, 1998; Rumberger, 2001; Wenger, 2002; Rumberger and Palaridy, 2005; Stutphen, 2010; De Witte and Cabus, 2011; among others). "Early school-leaving can be prevented through systematic evidence-based policies" (European Commission, 2011, p.5). Most dropout prevention measures "[…] do not attempt to change the institutional setting, but rather create alternative programs or institutions to target students who are somehow identified as at-risk of dropping

\(^1\) Following the European Commission, we define early school leavers as youngsters below age 23 who leave school without a higher secondary diploma.

\(^2\) In this paper high school and secondary education are synonyms.
out" (Rumberger, 2001, p. 23). The focus on at-risk students improves the cost-effectiveness of the prevention measures. This paper examines the importance of school absenteeism to identify students at-risk for school dropout.

Previous literature focused on the determinants of school absenteeism and its relation with school dropout in static model estimations (for an extensive overview, see Stutphen et al., 2010). They observe that truants and school dropouts are associated with similar student characteristics, such as male students, ethnic minority students, students from disadvantaged families, and low-ability students (Rumberger, 1983, 2004; Gary, 1996; Buscha and Conte, 2010). This paper takes an atypical stance as it estimates (1) the extent that unauthorized truancy accelerates the process of early school-leaving and (2) to which extent it may be considered a practical tool for policymakers and schools who wish to perform dropout prevention programs. This approach considers inevitable measurement errors in the nature and extent of the truancy spell.

We should make a precautionary note in that causal inference between truancy and school dropout is not straightforward. Both are influenced by unobserved variables such that endogeneity may arise. However, as truancy cannot follow from school dropout (by construction), we know the direction of the correlation such that a causal interpretation can be given. This is also argued in the data section 3 and empirical section 4 below.

**Duration models in education**

Students do not leave school overnight. The sequencing of events eventually leading to school dropout has been hypothesized in several theories of persistence (in the literature denoted by ‘student attrition’). It has been widely recognized that students are piling up problems before the actual dropout decision is taken (Rumberger, 2001; Psacharopoulos, 2007). Illustrative is the influence of the school, peers, and teachers (Spady, 1970; Tinto, 1975; Cabus, 2011), student motivation, ability and achievement (Cabrera et al., 1993; Sartarelli, 2011) and student intention and attitudes (Bean, 1980; Bean and Metzner, 1985; Price, 1977). Given the chronological nature of these events, this calls for a dynamic model. Duration models allow us to examine the timing of the dropout decision, as well as the increased hazard in dropout due to a precedent truancy record.

Duration models have only been occasionally applied in the educational literature (for example, see Singer and Willett, 1993; DesJardins et al., 1999, 2002; and Murtaugh et al., 1999). In addition, previous studies with duration models have mainly focussed on dropout at university level. To our best knowledge, there are 4 exceptions studying early school leaving. First, using a nonparametric single and competing risk hazard model, Bradley and Lenton

---

3In this paper, truancy and school absenteeism are used interchangeable. It denotes authorized and unauthorized school absenteeism, unless differently mentioned.
estimate the timing of the dropout decision among 16 to 18 years old voluntary enrolled in continued education in the UK (i.e., before entering university or higher education). The authors show that students drop out more frequently by the end of the study period (April, July). However, once they controlled for observable and unobservable characteristics, the probability of dropout is basically constant. Next, Light (1995) and Light and Strayer (2000) considered the re-enrollment decision of students, defined as the decision to go back to school after a dropout decision was taken. Using the semi-parametric proportional hazard model, they find that students re-enroll in school when costs are relatively low and benefits are relatively high, e.g., in times of an economic recession. Fourth, Olsen and Farkas (1989) discuss school dropout and childbearing outcomes using a continuous time hazard model. They estimate the so-called 'waiting time' to dropout of school and retrieve the determinants to retard this decision. For example, they provide evidence that the schooling level of the mother delays the dropout decision of the child.

**Contribution**

This paper contributes to the literature along three major lines. First, it focusses on school absenteeism and its influence on the timing of the dropout decision. This contrasts to previous conceptual and static models (starting from Tinto, 1975) whose adequacy has been questioned (Braxton et al., 1997). The timing of the event is estimated by a survival (or duration) analysis. This is convenient in the setting at hand as it creates a dynamic perspective and allows us to handle censored data arising from uncompleted spells. The paper will thus provide an answer to the question whether and to which extent truancy accelerates the dropout process.

Second, it accounts for data uncertainty. The main issue with previous literature arises from multiple definitions of truancy or self-reported measures. Very often, even within the same dataset, schools are using different definitions or interpretations of school absenteeism. This results in a lack of transparency, accuracy and reliability of the data, which makes the consistency of the outcomes intricate (e.g., Bos et al., 1992). This can be resolved by using an unequivocal definition of what is considered *unauthorized* school absenteeism. We therefore exploit an exceptionally rich administrative data set. Besides a uniform definition on authorized and unauthorized school absenteeism, the data reveal the measurement error for each spell: we observe whether the school considered a student as unauthorized truant, but in fact was not. This superior information is used to construct a Bayesian duration model where the measurement error is considered. It allows us to update prior beliefs on data validity into posterior beliefs conditional on observed data. As such, knowledge of the likely range of values of the parameters of interest are incorporated into the analysis. Accounting for Bayesian properties of data validity, we estimate uncertainty functions (or so-called *loss functions*).
functions; Koop et al., 2007), wherein the process of early school-leaving conditional on observed data is modelled.

Third, we analyze the relationship between truancy and school dropout in one particular European Union country, the Netherlands. The Netherlands is a particular interesting case study, as the Dutch Ministry of Education developed various measures to reduce dropout (for an extensive description, see: De Witte and Cabus, 2011). Among these various measures, improved reporting and registration provides us with a unique administrative dataset. As from 2005, every student who attends school in the Dutch educational system obtains a personal identification number, which is used to register truancy and school dropout. Similar information is only available in three other European Union countries, namely the United Kingdom, Germany and Italy (European Commission, 2011). Furthermore, there are uniform definitions in the Netherlands on what is considered (un)authorized truancy and school dropout. A school dropout is defined as a youngster who leaves school without a higher secondary degree between the age of 12-23. A truant is defined as a student who is not at school for one or more days. Within three days, the school reports the truancy to the ‘truancy officer’ who independently examines the nature and extent of the absence problem of the student. Only in the case of unauthorized truancy, the student ends up in the data with an unauthorized truancy spell.4 In case the spell was authorized, the data is corrected using a unique variable called report status. As such, the previous incidence of supposed truancy is still visible for the researcher. This is extremely useful in a Bayesian framework, where we exploit this superior information on authorized and unauthorized truancy. It provides feasible insights on the measurement error in the underlying data. In particular, we estimate a semi-parametric Cox proportional hazard model for school-aged students that appropriately accounts for the proportional hazard assumption and easily deals with Bayesian data properties. We show that only a Bayesian structure on the data yields insightful and consistent results.

The paper unfolds as follows. Section 2 outlines the Bayesian duration model in an education context. Section 3 discusses the data and descriptive statistics while Section 4 presents the results. A conclusion and discussion is offered in section 5.

---

4If the period of truancy was legitimated, the information is updated and the student will be no longer considered as truant.
2 Duration models for school dropout

2.1 Survival analysis

Denote by $T$ the time to an event (e.g., school dropout) and by $t$ the time under study. Further, denote by $t_0$ the time when an observation enters the analysis (e.g., day of birth). From $t_0$ onwards, all observations become at-risk of dropping out at school. We are interested in the likelihood that one does not dropout of school (i.e., likelihood of survival) and estimate:

$$ S(t) = 1 - F(t) = \Pr(T > t_0), $$

where $S(t)$ denotes the survival function and $F(t)$ the likelihood that an event occurs at time $T$, which is assumed to be a later time period than $t_0$ (van den Berg, 2000; Hjort, 2003). As students age, they face a higher risk of dropping out (Wenger, 2002; Entwisle et al., 2004, 2005; Plank et al., 2005). This is easily captured in the cumulative or integrated hazard function (Kalbfleisch, 1978b; Kalbfleisch and Prentice, 1973, 2002; van den Berg, 2000; Hjort, 2003):

$$ H(t) = \int_0^t \frac{f(u)}{S(u)} \, du $$

where the integrated hazard function $H(t)$ is equal to the natural logarithm of the survival probabilities or the accumulated age-specific failure rate. Equation (2) is equal to (1) when the survival distribution is continuous (Kalbfleisch, 1978b; Ibrahim et al., 2005).

Central in a survival analysis is the unobserved rate at which an event occurs, the so-called hazard rate $h(t)$. A hazard is defined as the age-specific failure rate and denotes, for observations who are under study (i.e., $T > t_0$), the (limiting) probability that the observation will fail a certain moment in time (i.e., in $(t, t + \Delta t)$) (van den Berg, 2000):

$$ h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t + \Delta t > T > t | T > t_0)}{\Delta t} $$

where $\Delta t$ is the interval between being at-risk and failure.

---

5 From a statistical perspective, we assume the time to an event as random.
6 The probability density function (pdf), $f(t)$ may be derived from $F(t)$ as follows:

$$ f(t) = \frac{dF(t)}{dt} = \frac{d}{dt}(1 - S(t)) = -S'(t) $$
As a student becomes at-risk of dropping out at school as from his/her birth onwards, the time under analysis \((t)\) may be expressed as years to the event since birth. This corresponds to:

\[
t = \frac{\text{date of dropout} - \text{date of birth}}{365.25}
\]  

(4)

Using equation (4), the output of the survival analysis can be expressed in age of the student at the time of dropout.

In the application below, we do not observe all students for three reasons. First, it is clear that for a duration model only observations with an eventual event are relevant. Second, due to data constraints, we consider only students who experienced an event in secondary education such that the data becomes left-censored (although the total number of subjects dropping out of primary education is expected to be very low; Wenger, 2002; Entwisle et al., 2004, 2005; Plank et al., 2005). Third, again due to data restrictions, we ignore school dropout above the age of 22 such that the data are right-censored as well. The left and right censoring provides an under bound estimate as the true effect will be higher or equal to the estimated effect.

To examine the relationship between truancy and school dropout, the survival probability function is made conditional on \((\text{truancy} \in \{0, 1\})\) and time \((t)\). This is denoted by \(S(\text{truancy}, t)\), where \(S(0)\) denotes the survival curve of regular school attendees and \(S(1)\) of truants. The division between truants and regular school attendees is retrospective: it has been assigned after failure. We proceed along three steps: (1) the nonparametric Kaplan-Meyer estimator to compute the survival probabilities, (2) a cox proportional hazard model to control for multiplicative effects and (3) a Baysian duration model to account for data validity issues.

**Step 1. Nonparametric Kaplan-Meyer estimator**

In a first step, we start from simply describing the data. The most simple way is to present the data in life tables (Cox, 1972). A life table estimates for each age the likelihood to drop out of school at that age (i.e., before the next birthday is reached). The additional probability to dropout of school due to precedent school absenteeism is then obtained by subtracting \(S(1)\) from \(S(0)\) at a given age:

\[
\Delta S(t) = S(0) - S(1).
\]  

(5)

The accumulated probability of dropping out at a given age is obtained by estimating the difference in the integrated hazard rate between truants \((H_1(t))\) and regular school attendees \((H_0(t))\). Therefore, we divide \(S(0)\) over \(S(1)\) and write:
\[ H_0(t) - H_1(t) = \ln \left( \frac{S(0)}{S(1)} \right) \]

Survival probabilities in a life table and the integrated hazard rates can be computed from the Kaplan-Meyer estimator (1958). The Kaplan-Meier estimator is the product of the probabilities that an observation survives at the end of a time interval given that it was present at the beginning of this time interval. This is a fully nonparametric estimation procedure and is in the literature also called the product limit estimator.

\[ S(t) = \prod_{t_i \leq t} \frac{n_i - r_i}{n_i} \]

where \( r_T \) denotes the number of failures at time \( T \) and \( n_T \) the number of survivors just before time \( T \). If time is discrete, a plot of the Kaplan-Meyer survival function is characterized by a step-wise function that visualizes failures (i.e., school dropouts). In the application below time is considered continuous such that the survival plot will be more smooth (i.e., it looks like a reversed S-function).

**Step 2. Multiplicative effects in a cox proportional hazard model**

To account for the influence of covariates, we rely on the semi-parametric Cox model (Cox, 1972). The model consists of two parts: (1) a non-parametrically estimated baseline hazard \( \lambda_0(t) \) which is assumed to be constant over time, and (2) a parametrically estimated set of covariates which have a multiplicative effect on the baseline hazard. The multiplicative effect is estimated by:

\[ \lambda_i(t) = \lambda_0(t) \exp(x_i' \beta) \]

where \( h_0(t) = \log_e h_0(t) \) denotes the unspecified base-line hazard at time \( t \), and \( \beta \) the vector of unobserved coefficients to the vector of characteristics \( x \) \( (x_1, ..., x_k) \) for observation \( i \) (with \( i = 1, ..., n \)).\(^7\) For a correct interpretation of the estimates, the baseline hazard is required to be constant over time. This is referred to as the proportional hazard (PH) assumption. If the baseline hazard is not a constant over time, it is argued that the event (i.e., school dropout) is associated with the analysis time. In other words, time would have an effect on school dropout. In our application, it is likely that school dropout is correlated with time (1) as older students are more likely to drop-out of school than younger students, and (2) as the compulsory education age is a clear cut-off between compulsory and voluntary education and

\(^7\)As the hazard is assumed to be positive, we specify the function of \( x \) as an exponential distribution. Indeed, an exponential distribution satisfies \( g(x) \geq 0 \) and \( g(0) = 1 \).
may alter dropout behavior of students. The validity of the proportional hazard assumption for the application is tested and discussed in Section 4. It is shown that the Cox proportional hazard assumption holds for students below age 18, such that dropout is independent of time for those students.

2.1.1 Step 3. A Bayesian duration model

As has been argued before, the administrative data set reveals the measurement error and data validity on the truancy variable. By a Bayesian structure, we exploit this superior information. Bayesian duration models are increasingly popular (e.g., Gamerman, 1991; Kalbfleisch, 1978; Kalbfleisch and Prentice, 2002; Kooper et al., 2007; Aitkin and Clayton, 2009). The Bayesian framework models start from the prior and posterior density on the data:

\[ p(\theta|y) = \frac{p(y|\theta) * p(\theta)}{p(y)} \]  \hspace{1cm} (8)

where \( y \) denotes the data at hand, \( \theta \sim (E(\Lambda(t)), \text{var}(\Lambda(t))) \) the population parameters, \( p(\theta) \) the prior density, \( p(\theta|y) \) is the posterior (observed) density, and \( p(y) \) is the marginal density of the observed data. The likelihood function \( p(y|\theta) \) is the likelihood function of interest: the probability that one may estimate \( \theta \) from the data \( y \). In other words, the Bayesian framework includes information (1) on the data \( y \) and (2) on the knowledge of the researcher with respect to the likely range of values of the population parameters \( \theta = (\beta, h, \sigma) \). For example, one may plot the right-hand side of equation (8) to visualize the shape of \( p(\theta|y) \) or one may use the data \( y \) to update beliefs on \( p(\theta) \).

Baysian structure in a non-parametric Kaplan-Meyer estimator  The Bayesian ideas can be included in the nonparametric Kaplan-Meyer estimator. At this point, it is insightful to briefly discuss the prior density used in a survival analysis. Koop et al. (2007, p.12) argue that, in case the prior density belongs to the exponential family, the prior density may be chosen so that the posterior density falls within the same elementary family of distributions as the prior. The (piecewise) exponential class of distributions (PE) have often been used in a nonparametric survival analysis (Gamerman, 1991). The latter partitions the analysis time in a finite number of \( K \) disjoint intervals, where random variable \( T \) follows a \( PE \) distribution \( T \sim PE(\lambda, \tau) \) with \( \lambda = (\lambda_1, ..., \lambda_k) \) and \( \tau = \{t_1, ..., t_{k-1}\} \). In each interval, one can estimate (Kalbfleisch, 1978):

\[ \text{Remark that the analysis time is ranked by } t_0 \leq t_1 \leq ... \leq t_{k-1}. \]
\[
\lambda(t) = \begin{cases} 
\lambda_1, t \in I_1 = [t_0, t_1] \\
\lambda_i, t \in I_i = [t_{i-1}, t_i], \quad 1 < i < k \\
\lambda_k, t \in I_k = [t_{k-1}, t_k]
\end{cases}
\]

where the hazard estimated for the \( i \)th interval or the probability to survive at time \( T \in [t_{i-1}, t_i] \) is conditional on \( T \geq t_{i-1} \). The intervals can be chosen as yearly intervals (i.e., \([t_{12}, t_{13}), [t_{13}, t_{14}), \ldots, [t_{21}, t_{22})\)). At the date of birth, there is no risk of school dropout such that \( \lambda(t_0) = 0 \). As from the age of 12 onwards (\( t = 12 \)), we may estimate the hazard as:

\[
\lambda(t_i) = \sum_{j=1}^{k} r_j \quad (i = 1, \ldots, n)
\]

where \( r \) denotes the cumulative risk of school dropout as from time \( (t_1 = 12) \) to \( (t_k = 22) \) (i.e., risk accumulated over all time intervals before event time \( T \)). It follows that \( r_{12}, \ldots, r_{22} \) are a posteriori independent and that \( \alpha_i = c^* \lambda(t_i) \) and a shape parameter \( c \). Correspondingly, the prior of the hazard follows a gamma process \( \Lambda \sim G(c \lambda, c) \), where \( \hat{\lambda} \) denotes the initial knowledge of the researcher on \( \mathbb{E}(\Lambda(t)) = \hat{\lambda}(t) \), \( \text{var}(\Lambda(t)) = \hat{\lambda}(t)/c \) and \( c \) the weight attached to that knowledge (Kalbfleisch, 1978, p.216). One may now easily see that the posterior distribution of \( \Lambda(t) \) is the sum of all independent risks over the intervals \( (j = 1, \ldots, k) \) (cf. increments) given that \( t_1 < t_2 < \ldots < t_k \) and for a particular \( c \).

If \( c \) is near 0, only little faith is attached to the initial knowledge of range of parameters \( \theta \). In this case, the likely range of parameters \( \theta \) is unknown such that a fully nonparametric method (e.g. the Kaplan-Meyer estimator) is consistent with \( c \). If \( (c \rightarrow \infty) \), the likely range of parameters \( \theta \) is assumed to be completely known. Here, one may find support of a fully parametric survival analysis (Kalbfleisch, 1978; Kalbfleisch and Prentice, 2002). We start in the data section 3 below with the assumption that \( c \) equals 0 such that a Kaplan-Meyer estimator is estimated.

**Bayesian structure in a semi-parametric Cox model** Under the Cox model (1976), the joint probability of survival of \( n \) subjects given the covariate matrix \( X \) is given by:

\[
P(T > t_0|\beta, X, \lambda_0) = \exp \left\{ -\sum_{j=1}^{k} \lambda_0(t) \exp(x_i' \beta) \right\}
\]

where the baseline hazard \( \lambda_0 = \hat{\lambda}(t_j) - \hat{\lambda}(t_{j-1}) \) is assumed to be the result of a stochastic process. The gamma process is often used as a prior for the cumulative baseline hazard function (Ibrahim et al., 2005).

\[
\lambda_0(t) \sim G(c \hat{\lambda}, c)
\]
The gamma process prior in equation (12) implies that $\hat{\Lambda}$ over the $j$th intervals are independent. However, in practice we only have to obtain a joint prior distribution for $(\beta, \lambda)$ as the baseline hazard is eliminated from equation (7) and enters the likelihood function by estimation of $(\lambda = \lambda_1, \lambda_2, ..., \lambda_j)$. For example, comparing two observations $i$ (e.g., a regular school attendee) and $j$ (e.g., a truant), we estimate the following likelihood function for $\beta$ (Kalbfleisch and Prentice, 2002; Ibrahim et al., 2005):

$$L(\beta, \lambda|\text{data}) = \prod_{i=1}^{n} \prod_{j=1}^{k} (\lambda_j \exp(x_i^j \beta))^\delta_{ij}v_i \times \exp\{-\delta_{ij}[\lambda_j(y_i - t_{j-1})] \sum_{g=1}^{j-1} \lambda_g(t_g - t_{g-1}) \exp(x_i^j \beta)\}$$

where $v = (v_1, v_2, ..., v_n)$ with $v_i = 1$ if the $i$th subject failed and 0 otherwise and $\delta_{ij} = 1$ if the $i$th subject failed or was censored in the $j$th interval and 0 otherwise. The observed data is denoted by $\text{data} = (n, y, X, v)$. A typical prior for $\beta$ is the normal distribution, where $\mu_0$ is the mean value and $\Sigma_0$ the variance. We then may write the joint posterior of $(\beta, \lambda)$ as (Kalbfleisch and Prentice, 2002; Ibrahim et al., 2005):

$$\pi(\beta, \lambda|D) \propto \prod_{j=1}^{k} [G_j \lambda_j^{(\alpha_0 - \alpha_{0,j-1})} - 1 \exp(-\alpha_0 \lambda_j)] \times \exp\{-\frac{1}{2}(\beta - \mu_0)\Sigma_0^{-1}(\beta - \mu_0)\}$$

3 Data and Descriptive Statistics

The application relies on a rich and unique administrative data set from the municipality of Amsterdam. The data consists of 12 to 22 years old students enrolled in secondary education in the Municipality of Amsterdam over the period 2005 to 2009. We count 13,538 unique students who dropped out of school in the period 2005 to 2009. This corresponds to about 5% of the data.\footnote{Note that some (although a minority of the students) may possibly return to school after being dropped out. This might be the case, e.g., in the subsequent school year after their first dropout decision.}

\footnote{Remark that $\dot{\lambda}(t) = (d/dt)\Lambda(t)$ and $\dot{\Lambda}(t) = \lambda t$ correspond to the unspecified baseline hazard from the Cox model.
\footnote{For further intuition on the estimate of $\beta$, one may consider:

$$\dot{\Lambda}(t|\text{truancy}) = 0 = P(T \geq t|\text{truancy} = 0, \lambda) = \lambda_0(t)$$
$$\dot{\Lambda}(t|\text{truancy}) = 1 = P(T \geq t|\text{truancy} = 1, \lambda) = \lambda_0(t) \exp(\beta\text{truancy})$$

where the relative risk ratio or hazard rate is equal to $\exp(\beta\text{truancy})$, the multiplicative effect of truancy on the baseline hazard and given that regular school attendees is the reference category.}
(following equation (4)), such that it equals to the age of a youngster at the day of dropout. In line with previous literature, a positive and significant correlation of 0.456 between age and school dropout is observed.\textsuperscript{12}

Descriptive statistics for the municipality of Amsterdam are summarized in Table 1. The dropout rate was 6.3\% in 2005-06 and decreased to 4.7\% in 2009-10 (Statistics Netherlands, 2011). The relative share of dropouts increases by age, with a major jump between the age of 17 and 18. The average dropout student leaves secondary education at age 18. Higher dropout rates are observed for men than for women, with a steady difference over the years of about 2 percentage points. Furthermore, we observe that ethnic minority students (in particular students from Aruba and Dutch Antilles) have relatively higher dropout rates than native Dutch students who have a dropout rate of around 4\%. Most dropout students are situated in vocational pathways or study subjects. The poverty measure (APCG) indicates if a student is living in a disadvantaged area which is characterized by lower incomes, more benefits receivers, and higher unemployment status (Statistics Netherlands). We observe a positive association between living in these disadvantaged neighborhoods and the school dropout rate with a multiplication of about 2 compared with residences out of the poverty area. Finally, in line with previous literature (e.g., Wenger, 2002; Entwisle et al., 2004, 2005; Plank et al., 2005), we observe in the bivariate analysis that large municipalities (>250,000 inhabitants) have higher dropout rates than medium (between 100,000 to 250,000 inhabitants) or little municipalities (below 100,000 inhabitants).

\textbf{A non-parametric visualization}

Using the nonparametric Kaplan-Meyer estimator (1958), we plot survival probabilities of regular school attendees and truants in Figure 1. The $Pr(T \geq t_{i-1})$ conditional on precedent truancy ($\in \{0, 1\}$) and years since birth is denoted by $\hat{S}(truancy; t)$.\textsuperscript{13} We count 7,254 (53.58\%) dropouts among regular school attendees and 6,284 (46.42\%) dropouts among the truants.\textsuperscript{14}

Figure 1 reveals a significant difference between truants and non-truants, although the two curves move parallel over the analysis time. This indicates that there are, depending on the truancy status, differences in the timing of the dropout decision. In general, lower survival probabilities for older secondary school students are intuitive and in line with pre-

\footnotesize
\begin{itemize}
    \item \textsuperscript{12} A truant is defined as a student who is not at school for one or more days. A school dropout is a youngster below the age of 23 without a higher secondary degree and who is no longer enrolled in secondary education (Dutch Ministry of Education).
    \item \textsuperscript{13} $S(truancy, t)$ is abbreviated by $S(0)$ for regular school attendees and $S(1)$ for truants throughout this paper.
    \item \textsuperscript{14} Remark that first failure denotes the first school dropout decision. As such, a small number of students reenroll in school after the first dropout decision was taken, however, a later moment in time they dropout one or more times in addition to the first failure.
\end{itemize}

\textsuperscript{12} A truant is defined as a student who is not at school for one or more days. A school dropout is a youngster below the age of 23 without a higher secondary degree and who is no longer enrolled in secondary education (Dutch Ministry of Education).

\textsuperscript{13} $S(truancy, t)$ is abbreviated by $S(0)$ for regular school attendees and $S(1)$ for truants throughout this paper.

\textsuperscript{14} Remark that first failure denotes the first school dropout decision. As such, a small number of students reenroll in school after the first dropout decision was taken, however, a later moment in time they dropout one or more times in addition to the first failure.
Table 1: Descriptive statistics of school dropout rates in the municipality of Amsterdam by
gender, age, ethnicity, poverty, residence, and vocational stream – Expressed as a procentual
share: the number of students in a given category to the total number of students in this
category.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total dropout rate</strong></td>
<td>6.3</td>
<td>5.7</td>
<td>5.5</td>
<td>4.4</td>
<td>4.7</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>men</td>
<td>7.5</td>
<td>6.8</td>
<td>6.8</td>
<td>5.4</td>
<td>5.7</td>
</tr>
<tr>
<td>women</td>
<td>5.2</td>
<td>4.5</td>
<td>4.3</td>
<td>3.4</td>
<td>3.7</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;13</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>14</td>
<td>1.4</td>
<td>1.2</td>
<td>1.1</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>15</td>
<td>1.8</td>
<td>1.5</td>
<td>1.3</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td>16</td>
<td>4.3</td>
<td>3.4</td>
<td>3.0</td>
<td>2.0</td>
<td>2.3</td>
</tr>
<tr>
<td>17</td>
<td>8.5</td>
<td>6.9</td>
<td>6.0</td>
<td>4.5</td>
<td>4.2</td>
</tr>
<tr>
<td>18</td>
<td>10.8</td>
<td>9.7</td>
<td>10.7</td>
<td>9.1</td>
<td>9.5</td>
</tr>
<tr>
<td>19</td>
<td>12.8</td>
<td>12.8</td>
<td>12.1</td>
<td>10.4</td>
<td>11.2</td>
</tr>
<tr>
<td>20</td>
<td>14.6</td>
<td>12.8</td>
<td>13.2</td>
<td>10.0</td>
<td>11.6</td>
</tr>
<tr>
<td>21</td>
<td>15.2</td>
<td>13.5</td>
<td>14.0</td>
<td>10.6</td>
<td>11.0</td>
</tr>
<tr>
<td>22</td>
<td>16.9</td>
<td>15.3</td>
<td>13.9</td>
<td>12.2</td>
<td>12.4</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native Dutch</td>
<td>4.8</td>
<td>4.5</td>
<td>4.3</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Suriname</td>
<td>9.5</td>
<td>8.2</td>
<td>8.4</td>
<td>6.9</td>
<td>7.4</td>
</tr>
<tr>
<td>Aruba/Antilles</td>
<td>8.7</td>
<td>8.6</td>
<td>10.5</td>
<td>7.8</td>
<td>8.7</td>
</tr>
<tr>
<td>Turkey</td>
<td>7.9</td>
<td>6.9</td>
<td>6.8</td>
<td>5.2</td>
<td>5.8</td>
</tr>
<tr>
<td>Morocoo</td>
<td>8.1</td>
<td>6.6</td>
<td>7.4</td>
<td>5.4</td>
<td>6.7</td>
</tr>
<tr>
<td>Other non-western</td>
<td>7.5</td>
<td>6.6</td>
<td>5.9</td>
<td>4.6</td>
<td>5.2</td>
</tr>
<tr>
<td>Other western</td>
<td>6.3</td>
<td>5.4</td>
<td>5.5</td>
<td>4.4</td>
<td>4.7</td>
</tr>
<tr>
<td>Unknown</td>
<td>26.9</td>
<td>28.4</td>
<td>22.6</td>
<td>23.7</td>
<td>22.3</td>
</tr>
<tr>
<td><strong>Poverty</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residence out of APCG</td>
<td>4.9</td>
<td>4.5</td>
<td>4.3</td>
<td>3.6</td>
<td>3.5</td>
</tr>
<tr>
<td>Residence in APCG</td>
<td>8.3</td>
<td>7.2</td>
<td>7.3</td>
<td>5.6</td>
<td>6.4</td>
</tr>
<tr>
<td><strong>Residence [inhabitants]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 20,000</td>
<td>4.1</td>
<td>3.7</td>
<td>3.7</td>
<td>3.1</td>
<td>2.9</td>
</tr>
<tr>
<td>[20,000;50,000]</td>
<td>5.0</td>
<td>4.7</td>
<td>4.3</td>
<td>3.3</td>
<td>3.6</td>
</tr>
<tr>
<td>[50,000;100,000]</td>
<td>4.8</td>
<td>4.9</td>
<td>4.4</td>
<td>3.6</td>
<td>3.5</td>
</tr>
<tr>
<td>[100,000;250,000]</td>
<td>5.2</td>
<td>4.7</td>
<td>4.6</td>
<td>4.0</td>
<td>3.6</td>
</tr>
<tr>
<td>&gt;250,000</td>
<td>7.8</td>
<td>6.7</td>
<td>6.8</td>
<td>5.3</td>
<td>6.1</td>
</tr>
<tr>
<td><strong>Vocational stream</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>without pre-vocational degree</td>
<td>2.3</td>
<td>2.0</td>
<td>1.9</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>with pre-vocational degree</td>
<td>10.5</td>
<td>8.4</td>
<td>6.7</td>
<td>4.9</td>
<td>5.3</td>
</tr>
<tr>
<td>vocational</td>
<td>14.3</td>
<td>13.1</td>
<td>13.0</td>
<td>10.8</td>
<td>11.5</td>
</tr>
</tbody>
</table>

Source: Central Bureau of Statistics Netherlands.
Figure 1: Survival probabilities of regular school attendees compared with truants

Previous observations. Nonetheless, there is a significant relative risk ratio. For example, the likelihood to survive at age 18 conditional on survival at age 18, $\hat{S}(0,18)$, amounts for regular school attendees to 56.9%. This likelihood drops to $\hat{S}(1,18) = 20.6\%$ for truants. Evidence indicates that more than one in five students with a truancy record leaves school before the compulsory education age is reached. Furthermore, the survival function shows that most of the truants leave secondary education between the age of 16 and 18 with a drop in survival probability from 72.5% to 20.6%. This is a remarkable difference to regular school attendees who have a survival probability at the age of 16 of 89.6%. These results provide us with a first indication that the process of school dropout is substantially accelerated when a youngster experienced a truancy spell.

**Bayesian methodology for data validity**

There are basically two sources of school absenteeism: authorized and unauthorized. Up to this point in the paper and similar to previous literature, we did not make a distinction between them. This ignorance, as well as using multiple definitions and self-reported measures, makes the analysis intricate since it is a source of uncertainty and measurement error (Bos et al., 1992). More precisely, the estimated integrated hazard for regular school attendees and truants $\left(\ln \frac{\hat{S}(0)}{\hat{S}(1)}\right)$ may be significantly different from $\left(\ln \frac{\hat{S}(0)}{\hat{S}(1)}\right)$ once accounted for misspecifications in the data. As outlined before, one can account for this by updating prior believes in the data structure (i.e., a Bayesian model).

To do so, we exploit an exceptional richness in the data: the administrative data indicates for every truancy spell whether (1) it was authorized (e.g., sick leave, wrongly reported), (2)
Table 2: Duration Models for school dropout by analysis time in years since birth

<table>
<thead>
<tr>
<th>Time</th>
<th>Duration Model 1 (standard)</th>
<th>Duration Model 2 (upper bound)</th>
<th>Duration Model 3 (lower bound)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S(0)\ln S(0)$</td>
<td>$S(0) - S(1)$</td>
<td>$S(0)\ln S(0)$</td>
</tr>
<tr>
<td>12</td>
<td>1.0042</td>
<td>0.0042</td>
<td>1.0187</td>
</tr>
<tr>
<td>13</td>
<td>1.0209</td>
<td>0.0207</td>
<td>1.0821</td>
</tr>
<tr>
<td>14</td>
<td>1.0586</td>
<td>0.0569</td>
<td>1.1808</td>
</tr>
<tr>
<td>15</td>
<td>1.1293</td>
<td>0.1216</td>
<td>1.3929</td>
</tr>
<tr>
<td>16</td>
<td>1.2361</td>
<td>0.2119</td>
<td>1.5455</td>
</tr>
<tr>
<td>17</td>
<td>1.6195</td>
<td>0.4821</td>
<td>2.0639</td>
</tr>
<tr>
<td>18</td>
<td>2.7596</td>
<td>1.0151</td>
<td>2.8648</td>
</tr>
<tr>
<td>19</td>
<td>5.0532</td>
<td>1.6200</td>
<td>6.1272</td>
</tr>
<tr>
<td>20</td>
<td>8.2625</td>
<td>2.1117</td>
<td>11.3597</td>
</tr>
</tbody>
</table>

unauthorized, or (3) still under investigation (i.e. final settlement on the nature of the truancy spell is missing). This variation in the settlement of the truancy spell allows us to obtain information on the data validity, which is used in the Bayesian framework to impose assumptions on the data. This contrasts the standard approach which assumes the data as 'clean' and without any error.

The data assumptions allow us to distinguish three models: a standard duration model, and two Bayesian models. First consider the standard model, denoted by Model 1. Model 1 is a standard duration model which assumes that the observed data are valid. In this sense, it follows previous literature. The results are summarized in Table 2. The first, second and third column provide, respectively, the ratio, the integrated hazard and the differences of two categories regular school attendees and truants at different moments in time (i.e., at different ages of the student). Whereas the likelihood to survive at a given age conditional on still being at school at this age (column (1)) steadily increases with age, a break can be observed at age 18. In particular, the likelihood is about three times higher at age 18 for regular school attendees than for truants. This was only 1.6 times higher at age 17, and about the same until age 14. Age 18 corresponds to the age when youngsters without retention in grade may legally leave high school (i.e., compulsory education age). The second column indicates the relative risk to dropout of school at a given age. Until age 17, this risk is relatively low while heavily increasing from age 18 on. For example, we observe a huge leap between the age of 17 ($\ln S(0)\ln S(0) = 0.4821$) and 18 ($\ln S(0)\ln S(0) = 1.0151$), indicating an increased proportional hazard or relative risk to dropout of school once the post-compulsory age has been reached (for an elaborated systematic review on retention in grade and its effect on at-risk students, see: Valentine et al., 2011). As can be observed from column (3), truants have a higher risk to drop out of school than non-truants. At age 15, a truant has a 15 percentage point higher risk to leave school than a non-truant. This increases to 17 percentage points at age 16, 30 percentage points at age 17 and further to 36 percentage points at age 18. After this age, we observe a decline in the relative risk.
Next a Bayesian structure is added to the data. We exploit superior information in the data on truancy. We distinguish two groups of truants. A first group are 'effective truants' as their truancy status has been confirmed by the truancy officer after an in depth research. Those students were illegally absent from school. A second group of truants consists of 'false truants'. Those students were reported by the school as truants, but in fact after research by the truancy officer, they had a valid reason to stay absence. The second and third model specification exploit the difference between the 'false truants' and 'effective truants'. The difference gives some insight in the data validity, and thus in the posteriori information structure on the data.

Model 2 estimates an upper bound to the estimates by considering only truants with an unconfirmed truancy status (i.e., false truants). Those students are reported by the school as truants, but in fact were not. In this sense, those students were falsely reported resulting in a too high truancy level. Focussing on those students gives an indication on the estimation bias as implying this information as posteriori beliefs in the Bayesian structure yields an upper bound estimate. The results in Table 2 indicate a considerable upward bias due to mis-specification in the data. The relative risk ratio shows that the values in column (1) and column (2) from Table 2 are systematically higher in Model 2 than in Model 1. This indicates that without the 'false truants', there is a higher dropout among truants than among regular school attendees. The 'false truants' are responsible for an underestimation of the influence of truancy on school dropout.

Model 3 estimates a lower bound estimation by considering only truants with a confirmed truancy status (i.e., effective truants). Those are all students for which we can be sure that they were without any reason or with an invalid reason absence from school. In this sense, it is a rigorous definition without data error. This results in a lower bound estimation. We observe that the lower bound in Model 3 approximately follows Model 1 before \( t = 18 \) with respect to the relative risk ratio.

Combining the lower and upper bound (Bayesian) estimations with the standard duration model delivers some insights on the validity of the data. Data are trustworthy if the outcomes of the standard model (Model 1) lie between the Baysian lower and upper bound (Model 2 and Model 3 in Table 2). Some clear differences are observed after age 18. In particular, the estimates of Model 1 are lower than the lower bound estimates of Model 3. This difference is due to the underlying data where from age 17 onwards more truancy spells are not finally settled. This indicates that the administration (in casu the truancy officer) puts less effort in these post-compulsory education age students. This is also presented in Figure 2, where by the age of 19 the average number of days that a truancy spell is under research doubled compared with age 16. In sum, the results of Model 1 are not falling between the lower and upper bound because of unresolved research: the nature of the truancy spell is unknown such
that it may falsely declare school absenteeism to be unauthorized. Given the illustrated data difficulties from age 17 onwards, the data are limited to school-aged students only.\textsuperscript{15} This finding has implications for our results provided in the following section 4 and is discussed next.

4 Results

Previous section argued to limit the data to age 12-17 and to update prior believes on the data by a Bayesian model. The remainder of the paper presents the results for the lower bound estimations (i.e., prior believes equal to unauthorized school leave by ‘effective truants’). Upper bound estimates are available upon request.\textsuperscript{16} Four different model specifications with increasingly additional multiplicative effects are estimated. The results are presented in Table 3.

\begin{table}
\centering
\begin{tabular}{|l|c|c|c|}
\hline
Model & Parameter & Estimate & p-value \\
\hline
Model 1 & Intercept & -2.57 & 0.032 \\
Model 2 & Intercept & -2.57 & 0.032 \\
Model 3 & Intercept & -2.57 & 0.032 \\
Model 4 & Intercept & -2.57 & 0.032 \\
\hline
\end{tabular}
\caption{Summary of result estimates.}
\end{table}

It has been argued in Section 2 that one needs to examine whether the proportional hazard (PH)-ratio is not violated. The proportional hazard test ($\text{Chi}–\text{square} = 0.32; \text{Prob} > \text{Chi}–\text{square}$) amounts to 0.570, supporting the crucial identification restriction of a constant baseline hazard in the limited sample (students equal or younger than age 18). The PH-ratio stresses the importance of the Bayesian modelling assumptions. If the uncertainty and measurement error in the underlying data would have been ignored, and thus a standard

\textsuperscript{15}From the Martingale residuals we observe further evidence in restricting the data to age 17. A plot of the estimated Martingale residuals is available upon request.

\textsuperscript{16}Note that the lower bound results (where information on ‘false truancy’ is exploited) are not presented here, as they are drawn on a smaller data set of n=413.
duration model is estimated, the PH-test \((Chi - square = 214.74; Prob > Chi - square)\) amounts to 0.000 such that the PH is violated. This is also reflected in the high estimated coefficient for \(\hat{\beta}\) of 2.233. Only a Bayesian structure on the data yields insightful and consistent results.

Model 1 is a basic Bayesian lower bound model which does not control for any source of heterogeneity. The estimate of interest is \(\hat{\beta}\): the multiplicative effect of unauthorized school absenteeism on the constant baseline hazard of dropout. The regular school attendees are thus the reference category. An estimated coefficient of 1 would indicate that regular school attendees and truants have an equal probability to dropout over the analysis time; an estimated coefficient larger (smaller) than 1 indicates a higher (lower) additionally probability to dropout during the analysis time. The estimate of \(\hat{\beta}\) in Model 1 equals 1.347 (significant at 1%-level) indicating that truants have a 34.7% higher likelihood to dropout of school compared to regular school attendees.

Cox Model 2 controls for individual, family and neighborhood characteristics. The estimated coefficients are in line with previous literature (e.g., Rumberger 1983, 2004). It has been argued in previous literature that students living in urban areas have a higher risk of dropout. We do not observe a similar correlation in the data as place of residence has only limited and insignificant impact on the dropout decision. Next, conditional on truancy status, we do not observe significant differences between male and female students. Ethnic minority students are considerably more likely to dropout of school compared to native Dutch students. Particularly other nonwestern and other western nationalities dropout more frequently. Students from single family households leave school significantly more without a higher secondary degree than students from two parents households.

In addition to the multiplicative variables, Model 3 includes school type characteristics. As the school type is strongly correlated to ability, students in different school types have different risks on early school leaving. For example, 12 years old students enrolled in supported education due to learning difficulties are more likely to be absent from school or dropout at a younger age than students who are enrolled in a pre-university pathway (Entwisle et al., 2004, 2005; Henry, 2007; Sutphen, 2010). Consequently, there is a higher relative risk ratio for lower ability education levels. The different school types (6 in total) are included in Model 3 by stratification. The stratification ensures that, even with varying subpopulations within an overall population, the proportional hazard assumption imposed on \(\hat{\beta}\) is not violated.

We use stratification based on school type additionally in Model 3. The Dutch education systems allows for 5 strata (Tieben and Wolbers, 2008): supported pre-vocational education (abbreviated with lwoo); pre-vocational education (vmbo); senior vocational education (mbo); general education (havo); and pre-university education (vwo). One may assume a unique baseline hazard to each specified stratum, however, we estimate equal coefficients.
across strata. The hazard ratio estimated in Model 3 drops to \( \hat{\beta} = 1.266 \) significant at 1%-level, indicating that the model overestimates the true impact of truancy if one does not account for (approximates of) ability.

In a final and fourth Cox model, we additionally control for (approximates of) ability. In particular, Model 4 limits the sample to only vocational students (lwoo, vmbo, mbo). This is in line with previous literature indicating that vocational students are most at risk to early school-leaving (e.g., Shavit and Müller, 1998; Gangl, 2002; Weltz, 2005). The total number of observations slightly drops from 4,203 to 3,331. The small decline in number of observations confirms that truancy is less of an issue in stronger educational tracks. Moreover, the estimated coefficient of the relative risk \( \hat{\beta} \) increases in comparison to the previous models to 1.374. This increase indicates that the relative risk ratio in higher ability levels is relatively small and lowers the estimated influence of truancy.

Overall, along the four model specifications, the estimate of \( \hat{\beta} \) is rather robust. Adding control variables does not significantly change the estimate which ranges from 1.26 in Model 3 to 1.37% in Model 4. Truants have a higher (accumulated) risk on school dropout by the age of 17 than regular school attendees. Note, however, that the underlying determinants on the truancy decision may be driven by particular exogenous covariates, e.g., by unobserved ability or motivation (for recent work on the determinants of unauthorized school absenteeism, see: Schaefer and Millman, 1981; Attwood and Croll, 2006; Henry, 2007; Sutphen, 2010). This calls for caution with a causal interpretation of estimates (Hausman and Woutersen, 2010).

This paper does not aim for a causal interpretation but rather explores unauthorized school absenteeism as a signal to identify at-risk students. It is desirable to include unobserved heterogeneity in this signal to obtain a good school dropout indicator. Furthermore, the direction of the effect of truancy on school dropout is known, as unauthorized truancy is a continuation of school dropout and the sign of the association is unambiguously positive. In sum, an unauthorized truancy decision leads to considerable higher probabilities to dropout earlier in the school career, but that the decision itself may be associated with student background.

5 Conclusion

This paper developed a nonparametric duration model to estimate the influence of school absenteeism on the risk of dropout in secondary education. Evidence indicates that eight in ten truants leaves school before the compulsory education age is reached, compared to only one in two for regular school attendees. In previous literature, school absenteeism is often a problematic indicator. Questionnaires are liable to significant measurement errors due to self-reporting. But even in administrative data, where this paper relies on, measurement
Table 3: Estimation results for regular school-aged students compared with students who were unauthorized absent from school (1)(2)

<table>
<thead>
<tr>
<th>Model</th>
<th>exp ( (x'_i - x'_j) \beta )</th>
<th>exp ( (x'_i - x'_j) \beta )</th>
<th>exp ( (x'_i - x'_j) \beta )</th>
<th>exp ( (x'_i - x'_j) \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(9.64)</td>
<td>(9.74)</td>
<td>(7.31)</td>
<td>(5.32)</td>
</tr>
<tr>
<td>1. gender</td>
<td>1.049</td>
<td>1.021</td>
<td>1.026</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.53)</td>
<td>(0.67)</td>
<td>(0.73)</td>
<td></td>
</tr>
<tr>
<td>2. ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suriname, Aruba</td>
<td>0.979</td>
<td>1.008</td>
<td>0.996</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-(0.47)</td>
<td>(0.18)</td>
<td>-(0.08)</td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td>1.084</td>
<td>1.081</td>
<td>1.153</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.44)</td>
<td>(1.37)</td>
<td>(2.27)</td>
<td></td>
</tr>
<tr>
<td>Marocco</td>
<td>1.093</td>
<td>1.106</td>
<td>1.174</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.83)</td>
<td>(2.05)</td>
<td>(2.96)</td>
<td></td>
</tr>
<tr>
<td>Other nonwestern</td>
<td>1.338</td>
<td>1.289</td>
<td>1.247</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5.83)</td>
<td>(5.03)</td>
<td>(3.82)</td>
<td></td>
</tr>
<tr>
<td>Other western</td>
<td>1.271</td>
<td>1.242</td>
<td>1.232</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.60)</td>
<td>(3.23)</td>
<td>(2.64)</td>
<td></td>
</tr>
<tr>
<td>3. family composition</td>
<td>0.794</td>
<td>0.811</td>
<td>0.794</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-(5.44)</td>
<td>-(4.93)</td>
<td>-(4.86)</td>
<td></td>
</tr>
<tr>
<td>4. residence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium municipalities</td>
<td></td>
<td>1.030</td>
<td>1.016</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.34)</td>
<td>(0.16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little municipalities</td>
<td></td>
<td>1.040</td>
<td>1.132</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.47)</td>
<td>(1.25)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Covariates: Individual, family, neighborhood characteristics
Specifications: Stratified, School type

Ties: Efron
No. subjects: 4,277
No. failures: 4,277
LR Chi-sq: 93.49
Prob > chi-sq: 0.000

(1) t-values between brackets.
(2) Only school-aged students 17 years old and younger are included in the analysis. With respect to truancy, only unauthorized school absenteeism is considered as defined and established by the compulsory education age officer. The analysis time corresponds to the age of the dropout students, as we imposed \( t = \frac{\text{date of dropout} - \text{date of birth}}{365.25} \).
(3) Only vocational students are included in Model 3 (i.e., lwoo, vmbo, mbo).
error may arise from the nature and extent of the reported school absenteeism. Using a rich administrative data set of the municipality of Amsterdam in this paper, we are able to consider this measurement error.

In particular, we improve data validity by relying on a Baysian duration model. By adapting the prior beliefs on the data, both lower and upper bounds on the survival probability are estimated. The results indicate that the data are inaccurate after the compulsory school-age of 17, suggesting a poor follow-up of post-compulsory school age students. The latter is likely the result of difficulties with defining the nature of the truancy spell once a student is no longer compelled to go to school, because of a lower sense of urgency to research, or a bad connection with the student and/or parents. This is undesirable: (1) for students who are prevented from further absence or eventual dropout by the compulsory education age officer, and (2) for policymakers and researchers who wish to evaluate school attendance, interventions and its implications for post-compulsory schooling.

Using a semi-parametric Cox regression with multiplicative effects, this paper compares the dropout decision of truants with regular school attendees. The results indicate that students with a precedent truancy record considered unauthorized leave school at an earlier moment in the school career than regular school attendees. Unauthorized truancy is less problematic in higher ability levels than in lower ability levels. For school-aged students enrolled in vocational education or training the dropout risk of truants compared with regular school attendees increases with 37.4 percent. The truancy decision leads therefore to a substantial accumulated school dropout risk by the age of 17, however, the decision itself may be driven by unobserved individual heterogeneity such as socioeconomic background (for more information on the determinants of school absenteeism and dropout, see: Rumberger, 1983, 2001, 2004; Henry, 2007; Stutphen, 2010).

In term of policy implications, this paper indicates that the identification of school-aged students with a precedent truancy record early in the school career can be a useful tool for dropout prevention. One straightforward way to use this tool as a prevention measure is to implement school attendance interventions or programs (i.e., sensitization of youngsters to attend class). Epstein and Sheldon (2002) and Sheldon (2007) argue that truancy rates can be influenced by preventive policies in elementary education. Helm and Burket (1989) and Roderick et al. (1997) provide evidence on the effectiveness of parental involvement on student attendance by telephone calls or timely information on student absences and school policies aimed at reducing truancy. Also class size composition and school policy may play a crucial role in student motivation and commitment (Spady, 1970; Tinto, 1975; Pittman, 1991; Lee and Burkam, 2003). In addition, various influences at the student’s place of residence or school may trigger both the dropout decision and truancy, such as a general dislike of school (Ehrenberg and Brewer, 1994), failing law enforcement or being part of discriminated
In connection with improved school attendance, one may provide at-risk students guidance and/or care through their study curriculum (Pittman, 1991; Phinney et al., 2006; Allen and Meng, 2010). Further research on the total number of days of the truancy spell may provide more insight in truancy as a dropout prevention measure. For example, not only the frequency of the truancy decision but also the duration of the truancy spell may be crucial for the school and/or teachers to undertake action.

References


